ORIGINAL

NATIONAL UNIVERSITY OF SINGAPORE

EE5907– PATTERN RECOGNITION

(Semester 2: AY2018/2019)

Time Allowed: 2.5 Hours

INSTRUCTIONS TO STUDENTS

1. Please write only your Student Number. Do not write your name.
2. This assessment paper contains **FOUR (4)** questions and comprises **FIVE (5)** printed pages.
3. Students are required to answer **ALL** questions.
4. Students should write the answers for each question on a new page.
5. This is a **CLOSED BOOK** assessment. One A4-size formula sheet is allowed.
6. Non-programmable calculators are allowed.
7. Total Marks is **ONE HUNDRED (100)**.

Q1 (25 marks). Subquestions (a) and (b) can be answered independently.

(a) Consider a binary classification problem of predicting binary class y from features x. The cost of correct prediction is \$0. There is a \$3 cost associated with predicting class 0 when the true class is 1. There is a \$6 cost associated with predicting class 1 when the true class is 0. Suppose the cost of asking a human to perform the manual classification is \$1. Therefore, for a particular x, there are three possible decisions: (1) decision α_0 predicts y to be 0, (2) decision α_1 predicts y to be 1 and (3) decision α_h requires a human to perform the manual classification. Let $p_1 = p(y = 1|x)$

(i) Assume the human is 100% accurate. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

(ii) Assume the human is only 90% accurate. Assume that when the human is wrong, the correct class is equally likely to be class 0 or class 1. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

(b) Consider a 2-class naive Bayes classifier with one binary feature and one Gaussian feature. More specifically, class label y follows a categorical distribution parametrized by π, i.e., $p(y = c) = \pi_c$. The first feature x_1 is binary and follows a Bernoulli distribution: $p(x_1|y = c) = \text{Bernoulli}(x_1|\theta_c)$. The second feature x_2 is univariate Gaussian: $p(x_2|y = c) = \mathcal{N}(x_2|\mu_c, \sigma_c^2)$. Let $\pi = [0.3 \;\; 0.7]$, $\theta = [0.7 \;\; 0.2]$, $\mu = [-1 \;\; 1]$ and $\sigma^2 = [1 \;\; 1]$.

(i) Compute $p(y|x_2 = 1)$. Note that result is a vector of length 2 that sums to 1.

(7 marks)

(ii) Compute $p(y|x_1 = 1, x_2 = 1)$.. Note that result is a vector of length 2 that sums to 1.

(6 marks)

Q2 (25 marks). Questions (c) can be answered independent of (a) and (b). Let $x \in \{0, 1\}$ denote the result of a coin toss ($x = 0$ for tails, $x = 1$ for heads). The coin is potentially biased, so that heads occur with probability θ_1. This information is reported through a noisy communication channel, resulting in outcome y. Because the communication channel is noisy, the result is correct with probability θ_2; i.e., $p(y|x, \theta_2)$ is given by:

	$y = 0$	$y = 1$
$x = 0$	θ_2	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

(a) Write down the joint probability distribution $p(x, y|\theta)$ as a 2×2 table, in terms of $\theta = (\theta_1, \theta_2)$. Note that $p(x, y|\theta) = p(y|x, \theta_2)p(x|\theta_1)$

(4 marks)

(b) Suppose we observe x and y eight times resulting in the following dataset:$\boldsymbol{x} = (1,1,0,1,1,0,0,0)$, $\boldsymbol{y} = (1,0,0,0,1,0,1,0)$, where the i-th element of $\boldsymbol{x}$ and i-th element of $\boldsymbol{y}$ correspond to the i-th data point. What is the maximum likelihood (ML) estimate of $\theta = (\theta_1, \theta_2)$? What is $p(D|\hat{\theta}, M_2)$ where M_2 denotes this 2-parameter model and $\hat{\theta}$ are the ML estimates? Please keep final answers in fractional form.

(11 marks)

(c) Now consider a model with 4 parameters, $\theta = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$ representing $p(x, y|\theta) = \theta_{x,y}$. Note that only 3 of the parameters are free to vary, since they must sum to one. What is the ML estimate of θ? What is $p(D|\hat{\theta}, M_4)$ where M_4 denotes this 4-parameter model and $\hat{\theta}$ are the ML estimates? Please keep final answers in fractional form.

(6 marks)

(d) Suppose we are not sure which model is correct. We compute the leave-one-out cross-validated log likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(m) = \sum_{i=1}^{n} \log p(x_i, y_i|m, \hat{\theta}(D_{-i}))$$

and $\hat{\theta}(D_{-i})$ denotes the MLE computed on D excluding the i-th datapoint. Which model will CV pick and why? Hint: notice how the table of counts changes when you omit each training case one at a time.

(4 marks)

Q3 (20 marks). Subquestions (a), (b) and (c) can be answered independently.

(a) Face photos captured for two persons with different expression are mixed together. It is required to design a face recognition model to separate faces of subject A from the ones of subject B. Two features are measured from the 5 photos for each subject and are provided as a training set given in the below.

Subject A		Subject B	
Distance between the right eye and nose	Distance between two eyes	Distance between the right eye and nose	Distance between two eyes
8.5	5.3	13.4	8.3
11.5	6.4	10.5	9.9
5.6	6.6	12.9	8.6
9.9	5.4	11.6	7.8
8.5	3.2	8.2	5.3

If only one feature from the above two can be used for the face photo classification, which one is better based on the Fisher discriminative criterion? Explain your choice in details.

(10 marks)

(b) K-means is one of the most popular clustering algorithm. List the objective function that k-means algorithm is optimizing in the iterative clustering process and list the main steps of k-means in a pseudo-code form.

(5 marks)

(c) Gaussian Mixture Model (GMM) can also generate clustering of data. Explain its difference from k-means clustering algorithm.

(5 marks)

Q4 (30 marks). Subquestions (a) and (b) can be answered independently.

(a) Suppose n training data $(x_i, y_i), i = 1, \ldots, n$ are provided, $x_i \in R^d$ is the feature of the i-th data sample, and $y_i \in \{-1, +1\}$ is the corresponding label. Let α_j denote the dual variable and $K(\cdot,\cdot)$ is a kernel function.

(i) Derive the following kernel SVM objective function from its primal objective form step by step.

$$\max_{\alpha} \; -\frac{1}{2}\sum_{j,k=1}^{n} \alpha_j \alpha_k y_j y_k K(x_i, x_k) + \sum_{j=1}^{n} \alpha_j.$$

(7 marks)

(ii) Suppose the data x_i is representing a graph-structured data (like molecular structure). The above question is changed to a graph data classification problem. Discuss whether kernel SVM model is still applicable to solve such graph-data classification. If yes, give a design for the kernel SVM model; if no, give the reason.

(8 marks)

(b) Consider a multi-layer neural networks and answer the following two sub-questions.

(i) Draw a $d - n_H - C$ fully connected three-layer neural network and label connection parameters between two adjacent layers on the network. Here denote the parameters as w_{ij} .

(5 marks)

(ii) Suppose the network is to be trained using the following loss function

$$L = \frac{1}{4}\sum_{k=1}^{n} (t_k - z_k)^4$$

Here z_k is a scalar and is the prediction for the input data $x_k \in R^d$ and t_k is the regression target for x_k. There are in total n data samples. Derive the learning rule Δw_{ij} for the hidden-to-output weights.

(10 marks)

END OF PAPER